SCIENCE AND TECHNOLOGY PEER REVIEW: GPRA

DR. RONALD N. KOSTOFF
OFFICE OF NAVAL RESEARCH
ARLINGTON, VA 22217
PHONE: 703-696-4198
FAX: 703-696-4274
INTERNET: kostofr@onr.navy.mil

(*The views expressed in this article are solely those of the author and do not represent the views of the Department of the Navy or any of its components.*)

ABSTRACT

This report describes practical issues for federal agencies to consider if they choose program peer review for internal purposes and/ or to contribute to satisfying the requirements of the Government Performance and Results Act (GPRA). For description purposes, the peer review process is divided into the following five phases:

1. Initiation of the review
2. Establishing the foundations for the review
3. Preparing for the review
4. Conducting the review
5. Post-review actions

Issues surrounding the various steps are presented in detail. Approaches are described for addressing issues that may arise during peer review. While the focus of this paper is on science and technology (S&T) peer review, issues considered and solutions to problems are applicable to other types of research. The Executive Summary (modified) from a much larger peer-review document is presented as an appendix, with some updates included.

KEYWORDS: peer review; GPRA; science and technology; strategic plans; metrics; performance goals; research evaluation; research assessment; technology assessment; evaluation criteria; research quality; research approach; research merit; review panel; investment strategy; research accomplishments; research transitions; journal

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **30 NOV 2003**	2. REPORT TYPE	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **SCIENCE AND TECHNOLOGY PEER REVIEW GPRA**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Ronald Kostoff;**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,800 N. Quincy St.,Arlington,VA,22217,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report describes practical issues for federal agencies to consider if they choose program peer review for internal purposes and/ or to contribute to satisfying the requirements of the Government Performance and Results Act (GPRA). For description purposes, the peer review process is divided into the following five phases: 1. Initiation of the review 2. Establishing the foundations for the review 3. Preparing for the review 4. Conducting the review 5. Post-review actions Issues surrounding the various steps are presented in detail. Approaches are described for addressing issues that may arise during peer review. While the focus of this paper is on science and technology (S&T) peer review, issues considered and solutions to problems are applicable to other types of research. The Executive Summary (modified) from a much larger peer-review document is presented as an appendix, with some updates included.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**37**	

publications; patents; research activity; research output; research impact; research outcomes; roadmaps; text mining; data mining; program review; project review.

INTRODUCTION

In 1993, the Government Performance and Results Act (GPRA, PL 103-62) was enacted into law (GPRA, 1993). GPRA applies to all federal outlay programs and has three components; strategic plans, annual performance plans, and metrics to show how well the annual plans are being met. Since the GPRA became law, there have been many federal interagency meetings to ascertain how the third requirement of the act, performance metrics, could be implemented to correctly portray the progress and accomplishments of S&T, especially research. There is a growing consensus in the S&T community that use of peer review is a more appropriate tool than metrics alone to measure S&T program performance in order to satisfy the GPRA requirements (Kostoff, 1997b). However, GPRA legislation states that if "it is not feasible to express the performance goals for a particular program activity in an objective, quantifiable, and measurable form, the Director of the Office of Management and Budget may authorize an alternative form" (GPRA, 1993).

The Office of Management and Budget, agreeing with the S&T community consensus, has authorized the use of peer review as an alternative GPRA metric for some agencies. Because of the expectation that at least some S&T sponsoring agencies will make use of this waiver, the volume of S&T program peer reviews across the federal agencies should increase dramatically. However, not only will the volume of program peer reviews change, but the conduct of the reviews will, of necessity, also change. If GPRA is fundamentally a budgetary instrument (Brown, 1996), then the performance evaluation results that contribute to the performance budgeting process must be of the highest quality. The methods chosen to obtain these performance evaluation results, program peer review and the supplementary quantitative performance measures, will require more rigorous and standardized operational characteristics than presently exist, such as process selection, reviewer selection, etc.

In 1997, a comprehensive document on research program peer review (Kostoff, 1997c) was placed on the Web. Its purpose was to bring the underlying issues and concerns surrounding research program peer review to the attention of the relevant research sponsoring, oversight, administering, managing, and performing communities. It was hoped that if these issues could be addressed comprehensively prior to full scale GPRA implementation, then procedures could be developed to conduct peer review in a manner that would not only support the performance budgeting process, but would also add value to the research programs being reviewed. To insure that the 1997 document reflected the experiences and findings of the larger research evaluation community, principles and findings from the manuscript and proposal peer-review literature were utilized, where applicable, to illuminate the research program review issues and help bridge the gaps in the research program review literature.

Since the 1997 peer-review document was placed on the Web, there have been a number of inquiries concerning very specific details of existing and proposed protocols contained within. The purpose of the present article is to expand on some of these detailed mechanisms that have proved necessary in the past for conducting an efficient and credible program peer review. The Executive Summary of the 1997 Web article (substantially modified) has been reproduced as an appendix to the present article. The reader interested in the broader topic of peer review is advised to consult the 1997 article, both because of its greater coverage and comprehensive list of references.

FIVE PHASES OF S&T PROGRAM PEER REVIEW

The S&T program peer review process can be divided chronologically into five somewhat independent phases. These are:

1. Initiation of the review
2. Establishing the foundations for the review
3. Preparing for the review
4. Conducting the review
5. Post-review actions

The following steps and considerations for each phase are recommended.

1. Initiation of the review
A successful S&T program peer review requires full participation by the unit undergoing review. Recalcitrance by the reviewee(s) can result in unacceptable delays, lack of necessary background information, and poor presentations. These deficiencies will hamper the review process and affect the quality of review results.

With few exceptions, no one likes or wants to be reviewed. How, then, can the unit undergoing peer review be motivated sufficiently to participate fully, and insure that the best review product will result? The author's experience from observing many different federal agencies' review processes is that motivation and participation derive from the actions of an organization's senior management at the initiation of the process. The management needs to communicate to thereviewees that they will be rewarded by appropriate participation and compliance in the review process, and penalized for non-compliance. Management needs to further communicate that critical judgments will be protected and handled with care. It is of the utmost importance that senior managementsend out an initial letter to all participants stating the following:

- The purpose of the review and its importance to the organization.
- The review's contribution to the larger agency GPRA response.
- The goals, objectives, and scope of the review.
- The identity and responsibilities of the review manager(s), the general responsibilities of the reviewees, and the responsibilities and reporting chain of the reviewers through all phases of the review process.
- The reviewees' performance both during the review development process and in the actual review will be part of their performance evaluation.
- The review manager will provide the input for thereviewees' performance during the review development process.

2. Establishing the foundations for the review:
Once the responsibilities have been assigned by the senior management, the principles that govern the review must be established. The review manager (ideally one person and not a committee) initiates this segment of the review by sending a letter to the senior management containing a detailed plan of how the total review process will be conducted. This letter is sent after extensive consultation on all review process aspects with the execution manager(s) of the unit(s) to be reviewed. Once this plan has been approved by senior management, the review manager sends a letter to the reviewees and all related support personnel, stating the following:

- The detailed objectives of the review.
- The process/approach to be followed in developing and conducting the review, including the evaluation criteria and the proposed disposition of the review report.
- A milestone schedule for completing all elements of the total review process, and
- assignment of personal responsibilities for completing each milestone.

The foundation elements to be discussed in detail in the plan, and in summary form in the reviewee letter, include the following items:

2.1. Identification of the boundaries of the program to be reviewed.

2.2. Establishment of a taxonomy that categorizes the program elements and defines the components by which the program will be reviewed.
2.3. Determination of the smallest unit (project, program) to be reviewed.
2.4. Identification of the evaluation criteria to be used.
2.5. Specification of the type of review group to be used (individual reviewer, fully independent panel).
2.6. Description of the different types of capabilities required by the review group (technical, managerial, application).
2.7. Identification of the types of attendees desired for the audience.

Considerations for each of these elements follow.

2.1. Identification of program boundaries
Identifying the scope of the program to be reviewed provides a framework for the remainder of the review. If the scope is defined too broadly (e.g., multiplepartially-related projects/ programs), then the review becomes very diffuse. This has consequences on the size and diversity of the panel required for a credible review. If the scope is defined too narrowly, the larger context and intrinsic integration and coordination with related projects may not be obvious. Unless thereexist hard bureaucratic boundaries and requirements that automatically set the review's scope, the scope definition phase should be iterated to achieve a balance between dilute focus and incomplete context.

2.2. Establishment of program taxonomy
The guiding principle for review options is that evaluation should occur along the same structures and taxonomies by which the S&T is planned and executed. If the agency has a separate S&T unit, then the technical area should be evaluated as an integrated whole. If research is vertically integrated with development, with concurrent planning and execution, then the research should be evaluated as part of a total vertical structure R&D review. A key conclusion to be drawn from this paragraph is that S&T evaluation recommendations must take into account how S&T is structured, integrated, and managed within an agency.

Establishing a taxonomy that represents the intrinsic nature of the program technically is analogous to selecting a mathematical coordinate system for solving a specific problem. Often, the ease of solving a particular technical problem, and sometimes the feasibility of solution, is highly dependent on selecting an appropriate coordinate system for the structure in question. This analogy holds for a program review as well. As in the mathematical system, the taxonomy selected should be orthogonal. This allows crisp presentations, each with a sharp focus, and minimal redundancy and overlap. Further, if the taxonomy contains too many categories, the review will be lengthened unnecessarily, and the program elements will appear to be discrete and fragmented. If the taxonomy has too few categories, it becomes very difficult to identify experts who can speak credibly for each component. Thus, a balance is required between selecting the appropriate number of review elements and ensuring that the review taxonomy remains aligned with the taxonomy used for program planning and execution. It has been the author's experience that time spent on the taxonomy definition phase results in time saved and problemseliminated downstream.

2.3. Determination of smallest review unit

Fiscally, an S&T, or research, program is a collection of funded S&T, or research, components. These elements could be subprograms, projects, or individual work units such as principal investigators (PIs). Conceptually, a program is greater than the sum of its components. A program includes the intelligence or inherent logic that links the components to each other and to the program's overall objectives. Thus, the intrinsic quality of an S&T or research program is not merely the sum of the qualities of the component projects, it depends on the quality of the structural relationships among and between the projects, as well as on the broader mission objectives.

Review of an S&T program can then be viewed as consisting of two elements:

2.3.1. "review of S&T projects," which examines the nature of the component projects, and is commonly referenced as an in-depth technical review; and
2.3.2. "review of an S&T program," which examines the nature of structural relationships among and between the projects and the mission objectives, and the relationships between the projects and the external environment.

This type of review is commonly referenced as a management review. These two elements, 2.3.1. and 2.3.2., can be merged operationally into a single review, or could be performed separately.

If review time were not a consideration, elements 2.3.1. and 2.3.2. would be recommended in total. This combination review would provide both depth and breadth necessary for a full understanding of program quality. In reality, review time is limited and it is desirable to have the same group of reviewers present for the total review of the areas in which they have expertise. This allows normalization and continuity to occur during the review action. However, in the case of a program review, the larger the program, the more review time it will require. It becomes more difficult to retain high quality reviewers as the length of the review increases.

There are at least three approaches to circumvent this problem. First, the program could be broken into focused subprograms, and each subprogram could be reviewed separately with more focused experts. Second, the program could have its components aggregated, and the full program could be reviewed by the same panel at a lower level of detail. Third, the quality and relevance components could be divided for separate reviews. While all the above options are theoretically possible, some compromise in quantity and type of material presented is necessary to insure that the same group of reviewers is presented with, and canevaluate, the totality of program material.

The author's experience and recommendations for GPRA are that a hybrid of elements 2.3.1. and 2.3.2. be presented. Since a program is being evaluated, it is important that the reviewers understand the total program's objectives, both in isolation and in the context of the larger organizational unit's objectives. It is equally important that the reviewers understand

- how the component projects relate to each other and the mission objectives,
- how they are integrated within the program and within the larger organizational unit, and
- how they are coordinated with the external environment.

At the same time, the reviewers should have substantial evidence that high quality S&T is being performed within the program. Thus, the review would centeraround the structural relations emphasis of element 2.3.2, with copious examples of technical progress and output and impact woven in the presentations where applicable. Notall technical details are required. Nevertheless, enough examples of positive accomplishments are necessary to convince reviewers of the effectiveness of the program. Because of the output/outcome/impact emphasis of GPRA, program reviews performed to partially satisfy GPRA requirements should focus on the S&T products and their potential or actual consequences.

2.4. Identification of evaluation criteria

Identification and selection of evaluation criteria should be driven primarily by the mission and review objectives, as well as the nature of material being reviewed. In the specific case of selecting evaluation criteria for peer reviews performed to address GPRA requirements, additional consideration must be given to selecting criteria of interest to the review client, as well as to the eventual disposition and utilization of the criteria ratings. If promoting the highest quality S&T to the relative exclusion of other objectives is the main program objective, then the evaluation criteria should focus on S&T quality. If accelerating transitions from research to development to demonstration is the prime program consideration, with S&T quality a secondary program objective, then the evaluation criteria should include both transitions and S&T quality, with greater weight given to transitions. If other program objectives are the main focus, such as integrating disadvantaged groups into the sponsored programs, then the criteria should included these goals and they should receive greater weight. In terms of the review mechanics, fewer criteria should be specified whenever possible. While it may be easier to analyze reviewer responses when many criteria are used, it forces the reviewers to fragment and channel their thinking and writing. The author has found that some of the most useful and coherent inputs are generated when the reviewers are allowed to provide comments in unstructured narrative form.

Reviews conducted by the author have allowed for a hybrid of both structured and unstructured types of inputs. For a research program, the fundamental evaluation criteria are:

- research quality,
- research relevance, and
- overall program quality.

The evaluation criteria recommended for a basic research review are addressed in the Executive Summary in the appendix. The criteria presented in the appendix resulted from separating research quality into its major components:

- research merit,
- research approach, and
- team quality.

For some evaluations, as shown in the full paper (Kostoff, 1997c), the fundamental evaluation criteria have been further subdivided into:

- research merit,
- research approach/plan/focus/coordination,
- match between resources and objectives,
- quality of research performers,
- probability of achieving research objectives,
- program productivity,
- potential impact on mission needs (research/technology/operations),

- probability of achieving potential impact on mission needs,
- potential for transition or utility, and
- overall program evaluation.

The full paper (Kostoff, 1997c) also presents sample evaluation criteria for more technology-oriented programs. Along these lines, a 2001 paper describes the review of an advanced technology development program in more detail (Kostoff et al, 2001b). If management or other non-technical issues are to be evaluated as part of the program review, then the evaluation criteria should be modified accordingly. Finally, the presenters should receive a copy of the evaluation criteria at the earliest stages, so that they can begin to craft their presentations to focus on addressing the criteria.

2.5. Review group type

Selection of the type of review group is a core issue, and should be addressed at the initiation of the review process. While many types of groups are possible, two will be discussed here. They are the independent panel (2.5.1) and the external reviewers group (2.5.2).

2.5.1. Independent panel.

The independent panel is a group of experts independent of the agency, and typically funded under a contract. The independent panel has a chairperson, attempts to reach consensus on issues, and generates a written report containing the results of the review and sometimes recommendations.

2.5.2. External reviewers group

The group of external reviewers consists of experts individually contracted to the agency. The reviewers report to the agency review manager. The external reviewers group does not have a chairperson; the review manager serves this role. While the group may engage in technical discussions during the course of the review, it does not reach a consensus. While there may be individual written inputs from each group member, there is no group report. The review report is written by the agency review manager based on the individual written inputs plus other considerations. Because of the technical understanding required to write a credible report, as well as select the appropriate mix of reviewers, and conduct all aspects of the review, the review manager should have a solid technical background and some understanding of the subject matter to be reviewed.

Each of the two review group approaches has value for specific applications. The group of external reviewers is less formal, and has fewer reviewer and audience restrictions. It is useful for internal reviews where structural program issues are paramount and need resolution or improvement, and where comparison with other programs is not the major focus. The independent panel is more formal. The independent reviewer panel has more specific reviewer, meeting, and audience selection constraints/requirements. If the panel is run under the auspices of one of the National Academy of Sciences boards, for example, there will be a more elaborate process used to select participants and review the final written product. From the agency's perspective, either group has very high utility for addressing the agency's program improvement needs. From a perspective external to the agency, the independent panel has higher credibility because of its independent nature. For GPRA application, the independent panel is more appropriate, because of its perceived independence.

However, operation of an independent panel under GPRA will be intrinsically different from past operation of this type of panel. If GPRA is viewed as a budgetary instrument with a potential for modifying resources (Brown, 1996), some additional factors must be considered in structuring and operating the two types of panels discussed. Since different types of panels may be used for different technical areas and different agencies, some means of normalizing review results across areas and agencies will be required. Also, because of the potential for errors or bias, some means of rebuttal or reclama must be provided for conclusions and recommendations produced by different panel types. Both these issues are summarized below.

2.5.3. Review report normalization

The author has not seen any fully satisfactory peer review normalization approaches due to the presence of many non-separable variables. However, one interesting normalization approach is used by the Dutch Technological Foundation for evaluating research proposals (Van den Beemt, 1991, 1997). Technical comments, but not quality ratings, are provided by technical peers. The comments and proposer responses for twenty different proposals are then provided to twelve people from a variety of disciplines. This "jury" of twelve provides the scores through an independent mail review. Essentially, the normalization is provided by having the twelve jurors common to all proposals.

The author has used two approaches to improve normalization across panels somewhat. First is the utilization of some individuals common to all panels. In a series of competitions for new accelerated research programs that was held in the late 1980s (Kostoff, 1988), the author served as de facto chairperson of all the different discipline panels. This resulted in some small measure of normalization among the different panels. Use of more individuals common to all panels would have provided an extra measure of normalization, and in this sense the presence of senior management during the reviews provided additional measures of normalization.

Obviously, the more closely the panels are related topically, the more valuable is the technical contribution of individuals common to the different panels. Secondly, in the above competitions, it was assumed that the difference in aggregated average scores for major disciplines (e.g., physical sciences and life sciences) was due to two factors: differences in intrinsic quality of the programs proposed and differences in the scoring severity of the reviewers. To normalize, a fraction of the differences in aggregated average scores for the major disciplines was removed. This was assumed to eliminate the scoring severity difference. Trial and error showed a fifty percent correction factor provided results that appeared reasonable to the audience members who had attended all the reviews. This normalization procedure had the added benefit of preserving and insuring representation from disciplines that had strategic value to the organization. This approach to normalization could have a second interpretation. If the research is viewed as having a strategic component and a quality component, with the reviewers' scores viewed as addressing the quality component only, the correction could be perceived as adjusting for the presence of the strategic component.

For example, assume a life sciences panel produced an average program score of five, and an engineering sciences panel produced an average score of ten. Assume further that each discipline had equal strategic value to the organization and that the strategic value (STRAT) was perceived by the organization to be of equal importance to the reviewers' scores (SCORE-assumed to be a total program quality score that includes mission relevance). Then the normalized total score (FOM) can be computed as FOM = 0.5*STRAT + 0.5*SCORE, and the difference between the two panels' scores would be reduced from five to 2.5. This correction factor can then be applied to the raw score of each program within the discipline to arrive at a final "normalized" score.

2.5.4. Rebuttal of review panel recommendations

In a 1997 paper (Armstrong, 1997), different studies of errors and superficial work by peer reviewers of journal manuscripts are described. The conclusion one draws from these results is that the problem of manuscript reviewer error production is not insignificant. In most research program peer reviews, commission of technical errors by reviewers due to the relaxed standards resulting from anonymity and lack of financial incentives is probably not nearly as serious as in manuscript reviews. In the author's experience, panel members tend to suppress overt expressions of biases, and they typically make statements they are able to defend. Studies of the extent of errors, or bias, committed by research program peer reviewers remain to be done. If these panels eventually have substantial input to the budgetary process under GPRA, an appeals system for program reviews may have to be established to resolve errors or perceived biases.

2.6. Specification of review group capabilities required

Even with the strongest support from an organization's top management, and the direction of an unbiased and competent review leader, the quality of a review will never go beyond the competence of the reviewers. Two dimensions of competence that should be considered for a program peer review are the individual reviewer's technical competence for the subject area, and the competence of the review group as a body to cover the different facets of S&T issues (research impacts, technology and mission considerations and impacts, infrastructure, political and social impacts). The quality of a review is limited by the biases and conflicts of the reviewers. The biases and conflicts of the reviewers selected should be known as well as possible to the leader and among the reviewers themselves.

One common error in panel selection is limiting the choice of S&T experts to those who have specific expertise in the subdisciplines of the existing program. This provides an answer to the question of whether the job is being done right, but not whether the right job is being done. The former question relates to detailed technical quality, while the latter relates more to investment strategy in the broadest sense (investment strategy is the rationale for the prioritization and allocation of resources among the program components.). To answer the latter question, people with broad expertise in the area covered by the overall program's highest level objectives should also be selected. They will be able to address the investment strategy more objectively, and determine whether the mix of subdisciplines and the allocation of resources among the subdisciplines is appropriate. The review group, then, would be able to address the central question of whether the right job is being done right.

One of the major criticisms of peer review, whether manuscript, proposal, or program, is that it tends to perpetuate orthodox and conservative paradigms, and tends to reject new paradigms that threaten the structure of the status quo. If one of the objectives of an S&T program peer review is in fact to ensure that innovation is recognized, that truly revolutionary research with attendant new paradigms will be promoted and rewarded, then the selection of reviewers to address the right job issue in parallel with reviewers to address the job right issue becomes of paramount importance.

In summary, a review panel should have at least the following characteristics:

- Each member should be highly competent in the facet of the program for which he/she has been selected; this assures the presence of sufficient depth on the panel.
- The panel as a body should have sufficient competence to cover all major facets of the program being reviewed; this assures the presence of sufficient breadth on the panel.
- Each member should be minimally conflicted with the program under review, and any conflicts or biases should be known to all the panel members before the review; this assures the presence of independence and objectivity on the panel.
- Each member should agree to read all background material, attend all sessions, and protect any classified and proprietary information that surfaces during the review; this assures the presence of preparedness and security on the panel.

2.7. Identification of audience types

A program review provides an excellent forum for disseminating program information and results to a wide audience. In addition, a program review is a useful mechanism for providing coordination with intra- and inter-organization related programs. Care should be taken to insure that the review audience includes:

- actual and potential customers,
- stakeholders and other oversight groups,
- co-sponsors,
- users, and
- other agency representatives.

Judicious use of the many databases that are now accessible, and algorithms that expand the identification of potentially related technical areas and their contact points (Kostoff, 1997e, 1999b, 2000a, 2001c, 2001d, 2003a, 2003b, 2003c, 2003d) can help develop a broadly-based audience for maximum impact.

3. Preparing for the review

The schedule and milestones originally submitted to senior management to obtain approval for initiating the review should be further detailed. A tracking system for schedule progress should be initiated and periodic status reports sent to senior management. The author has found weekly status reports to be adequate.

3.1. Developing the agenda

Once the taxonomy has been developed, the structural elements of the agenda can be easily identified. The main elements include:

- an introduction by the review manager to identify the goals of the review, set the stage for the remainder of the review, and handle any administrative issues;
- an overview by the program manager of:

  - the role of the program in its larger context,
  - the vision of the operational scenario to which the program will contribute,
  - the requirements necessary for the vision to be achieved,
  - the technical capabilities defined by the requirements and the S&T necessary to produce the capabilities,
  - promising S&T opportunities that could result in capabilities not yet defined by requirements,
  - the overall investment strategy that links the above components to each other and to the external environment and will allow the capabilities to be obtained, and
  - the detailed technical presentations to follow.

- detailed technical presentations and, if these are held at a laboratory, tours could be included in this segment;
- question and answer time allocated to each presentation;
- written evaluation periods after each presentation;
- an executive discussion period at the end of each day; and
- administrative break periods (coffee, lunch, etc.).

3.2. Developing the presentations

3.2.1. Assignment of responsibilities

The presentation development phase begins by assigning the responsibility for the presentations to the program manager. The program manager is sent a letter detailing these responsibilities, identifying:

- overall time available on the agenda for presentations,
- fraction of presentation time reserved for questions and answers,
- taxonomy to be used for evaluating the program, and
- criteria by which the program will be evaluated.

The program manager then has to decide:

- the amount of time to be devoted to addressing each taxonomy category,
- how to address the category, and
- who should make the presentations for each category.

There is a wide range of combinations of potential presenters for the total program being reviewed. At one extreme, the total program presentation could be made by the program manager alone. At the other extreme, each taxonomy category could be presented by selected PIs (the performers). The level of presenter selected depends on the objectives, type, and location of the review. For a GPRA-type program review conducted at a sponsor's headquarters, the author's preference would be to have as few different presenters as is feasible. Each presenter should be as high in the program management chain as possible while still having an acceptable grasp of the technical material. This allows the program integration message to be communicated to the audience most effectively. For a smaller program review conducted at a laboratory, in which tours of the working environment may be incorporated, PI-level presentations could be included.

3.2.2. Reducing presentation problems

The reasoning behind recommending that presenters be relatively high in the program management chain is the following. For the large federal S&T sponsoring agencies with which the author is familiar, technical competence of the performers is not a major issue or problem. The number of proposals to these agencies far exceeds the funding available, and with the use of in-house and external experts to provide advice in proposal selection, typically only the 'cream-of-the-crop' is selected. Reviews in which the author has participated that focus mainly on technical quality at the PI level invariably arrive at the conclusion that the technical work is of high quality. This conclusion appears almost invariant of the agency or type of panel or reviewer selection process employed. If a problem is surfaced, it tends to focus on the following issues of integration and coordination:

- Are the different projects coordinated with each other and with other agency projects?
- Do they form a cohesive program or are they a collection of isolated and fragmented efforts?
- Are the projects coordinated/jointly planned/jointly managed with external organizations and is the total program coordinated in this way with the external community?

The actual S&T performers tend to focus on the technical details, and the coordination and integration issues are best addressed by those somewhat removed from the actual performance of the tasks.

Another presentation problem that appears to emerge in every agency presentation the author has attended overlaps somewhat with the technical detail/coordination issue described above. The problem stems from the training and characteristics of many S&T performers. Technical personnel are trained to pay careful attention to details, and very good technical people seem to have an innate interest and predilection for details. While some technical presentation skills are included in technical training, they typically constitute a small portion of that training. Consequently, many program level presentations remain immersed in technical details and tend to be far too long. While this level of presentation is most comfortable for the technical specialist making the presentation, it acts to the detriment of presenting the program in its larger context. In addition, because of the concentration on details, the main message tends to become diluted and diffuse and overwhelmed by material extraneous to the main message. It is very important that the main message to be delivered be kept in focus at all times when structuring the presentations. More specifically, the presentations should be kept short and the number of view graphs should be few. Every line (and word) on each view graph should contribute to the central message that the presenter wants to communicate. If it does not, it should be removed. The producers of TV commercials have learned this lesson well. Unfortunately, these fundamental communication principles and techniques have not found their way to many technical program presenters.

3.2.3. Presentation content

3.2.3.1. Outline of presentations

In alignment with the agenda outline, the detailed contents of the specific presentations should incorporate the following. There should be an overview showing how the larger management unit (division, department, etc.) in which the programs are housed integrates into the total organization, and how the management unit's objectives relate to those of the larger organization. Then, the investment strategy of the larger management unit should be presented in detail. The investment strategy presentation should include the:

- relative program priorities,
- actual investment allocation to the different programs, and
- rationale for the investment allocation.

Finally, for each program presentation, the investment strategy for its thrust areas should be presented. The investment strategy is perhaps the most crucial part of a program review, and deserves further discussion here.

Investment is the allocation of resources among the program components. Investment strategy is the rationale for the prioritization and allocation of resources among the program components. The optimal investment strategy for a program is the specific allocation and rationale that will produce the most mission relevant high quality S&T for impacting the program's objectives. This will depend on the viewpoint of the assessor and, in particular, how the assessor limits the role of the S&T within the national perspective.

The optimal investment strategy should be a focal point of an assessment. The optimal investment strategy results from a timely confluence of:

- S&T requirements (top-down driven) and
- promising S&T opportunities (bottom-up driven).

Further, promising S&T opportunities result from a timely confluence of advances in:

- theory,
- instrumentation,
- new experiments,
- new algorithms, and
- computers.

Finally, S&T requirements result from a timely confluence of:

- domestic and foreign,
- political and economic, and
- strategic and tactical advances.

All of the above factors should be included in a presentation of the investment strategy.

3.2.3.2. Specific presentation content

The senior management presentation.

To initiate the actual review, a senior agency manager provides a short introduction describing structure and mission of the agency, and a more detailed description of the purpose and goals of the program review. Senior management describes what is expected from the reviewers, and how their comments have been, and will be, utilized.

The review manager presentation

The review manager provides the details of the organization's structure, the types of reviews within the agency, and the integration of the present review with the other reviews and with the total organization's management processes. The review manager also describes the steps of the specific evaluation process, including the meeting agenda, and presents all the administrative details and procedures to be followed.

Organizational unit head presentation

The broader technical portion of the presentations is initiated by the head of the organizational unit in which the program resides, and it includes the following informational material:

- The mission and objectives of organizational unit,
- a list of all programs in organizational unit,
- a description of objectives of each program,
- the funds and people associated with each program and with the program to be reviewed,
- an overview of the accomplishments and transitions of programs not being reviewed, and their relation to the accomplishments and transitions of the organizational unit's mission and potential national impact, and
- responses to actions taken as a result of the previous year's reviews of the organizational unit's programs

Program manager presentation

The program manager(s) then provides a more detailed overview of the program under review, including:

- objectives of program under review.
- requirements to be met and derived target capabilities for the S&T initiative (For example, in the review of a military-oriented program, what is the present and evolving threat-identify documented sources, personal contact sources, etc.?

What is the importance of the threat and what are the capabilities required to overcome the threat?).

- investment strategy.
- list of targeted thrust areas selected to meet program requirements (e.g., propulsion, aerodynamics, G&C) and sub-thrusts (e.g., energetic propellants, combustion instability, propellant safety).
- objectives of each thrust that will include:

  - thrust and sub-thrust funding and prioritization,
  - rationale for thrust and sub-thrust selection and prioritization (including the bases for rationale and prioritization such as system studies, workshops, assessments, intuition, Congressional and other mandates, etc.),
  - integration of thrusts and sub-thrusts to form overall program coordination/roadmaps (Road maps are graphical displays of the inter-connectivity among diverse S&T projects and potential applications. They describe the past, present, and future of the program, and its linkage to other internal and external programs, as well as linkage to institutional capabilities and requirements. They offer a convenient focal point for discussing complementary and related programs sponsored by other external organizations.),
  - team quality (identify S&T performers), and
  - a summary of major accomplishments, transitions, milestones met.

The technical manager presentation.
The technical managers who support the program manager will present the following:

- Objectives of each sub-thrust
- Technical roadblocks to achieving the sub-thrust objectives
- Technical approach for overcoming the sub-thrust roadblocks
- Potential sub-thrust payoffs and capability enhancements
- Technical results achieved

3.2.4. Dry runs
After the presentations have been developed and reviewed within the performer organizations, there should be at least two series of "dry runs" before the review manager. If possible, senior management should be in attendance as well. The dry run presentations should be polished from the presenter viewpoint, and the main purpose is to assure that all the separate taxonomy category presentations appear

cohesive and integrated. The dry runs are not forums in which diplomacy and tact, and the preservation of fragileegos, are paramount. One key objective is that all questions and issues and weak points that could arise in the final presentations are surfaced and discussed in the dry runs. The earlier such issues are resolved, or at least recognized, the better for all participants.

3.3. Selecting and inviting the reviewers

Selection of an optimal review panel is more of an art than a science, and depends on:

- the selector's understanding of the many facets of the program being reviewed,
- his/her understanding of the experts available in the technical community, and
- his/her ability to predict the interaction dynamics of a particular group of experts.

Presently, different federal agency approaches in panel selection range from assembling program manager recommendations as potential reviewers to using an iterative co-nomination approach for reviewer identification and selection. Since the latter approach, properly done, is relatively objective to the program being reviewed, it will be the focus of this discussion.

In essence, the iterative co-nomination approach is a multi-step process that starts with an input list of recommended experts and results in a list of experts who have been multiply nominated by different experts. Once the overall technical description of the program is generated, and technical descriptions of the taxonomy categories (technical sub-areas) are provided, reviewer identification can be initiated. Sources of candidate reviewers can include:

- program manager recommendations,
- membership lists of prestigious organizations such as the National Academies of Science and Engineering and the Institute of Medicine,
- agency review boards,
- agency consultant pools,
- contributors to technical databases (such as journal article authors or technical report authors), and
- other similar lists.

Multiple names are chosen to cover:

- each sub-discipline,

- the program as a whole,
- allied research disciplines,
- the technologies, systems, and operations that the program does or could potentially impact, and
- other elements of the customer, stakeholder, user, and impactee communities.

This list of names is called level 1, or the initial list. Each member of level 1 is asked to identify, or nominate, other experts in his/her particular area of expertise to generate the level 2 list. For example, assume that a physics program is being assessed. Assume further that this program has three subdisciplines: plasma physics, atomic physics, and molecular physics. The level 1 list may have two names for each one of the subdisciplines. To obtain the level 2 list for the plasma physics research area of expertise, each of the two plasma physics recommendees of level 1 would be asked to recommend two experts in plasma physics. If names appear more than once in the level 2 list, or between the level 1 and level 2 lists (multiply recommended individuals), then these individuals are assumed to be the leading experts in the fields to be assessed. If no multiple recommendations appear, then the experts in level 2 are asked to recommend two experts in plasma physics for level 3, and the co-nomination search is repeated. Convergence occurs when an adequate number of experts have been co-nominated. While this process may at first seem complex and open-ended, convergence is rapid because of the relatively small number of real experts in any well-defined technical discipline.

A primary and alternate list of co-nominees should be matrixed against selection requirements and criteria, where the matrix elements represent the reviewer's expertise in the different facets being examined. This matrix should be distributed to the program managers and performers who will be reviewed, and comments related to bias and conflict solicited. If strong objections can be supported against one or more nominees, the list could be modified. Some additional constraints should be placed on the list of reviewer candidates. Because the iterative co-nomination approach focuses on identifying recognized experts in a field, there is always the danger of excluding younger reviewers of high caliber with fresh perspectives on the topical area. Therefore, the co-nomination approach has to be tempered with other selection processes that allow for the recognition of lesser known experts of high quality.

In practice, the author uses a hybrid combination of reviewer sources and selection approaches to insure that a diversified portfolio of appropriate experts is represented on the review team. There needs to be a balance of continuity and turnover among reviewers. The ratio between these two considerations will be heavily dependent on review frequency. For three year period reviews, the author has tended to use about 25-33% continuity. Total number of reviewers is another important consideration. As the number of reviewers on the panel increases, more coverage of depth and breadth is possible, and the diversity of opinion on a given topic area is increased. At the same time, the cost of conducting the review increases, and the logistics of controlling the panel increases. The author has found that a range of panel sizes from about eight to fourteen is desirable, with the actual size depending on the range of material covered by the review. Once the list has been finalized incorporating the above considerations and constraints, potential candidates are contacted by phone. If there are no conflicts-of-interest, invitations are then extended, preferably at least three months in advance of the review date.

3.4. Selecting and inviting the audience

As stated earlier, care should be taken to insure that the review audience includes actual and potential customers, stakeholders and other oversight groups, co-sponsors, users, impactees, and other agency representatives. The invitation may come from the program manager(s). Databases, however, can help in the identification of other participants. Depending on how the GPRA reviews are conducted, especially who is conducting them and where they are being conducted, announcements to the general public may be advertised. While a large audience in a review room may serve to restrict discussion, with the present-day ease of establishing video transmissions, separate rooms can be reserved for general public audiences remote from the review room. Once the desired audience has been identified, invitations should be sent at least three months in advance of the review. This substantial advance notice will insure that the busy schedules of high caliber attendees can accommodate the review. The invitation package should include many of the elements sent to the reviewers, including the background material.

3.5. Selecting and distributing background material

It is strongly recommended that a variety of background material be supplied to the reviewers (and the invited audience) before the review. This should include:

- material focused strictly on the internal program under review,
- material focused on related external programs, and
- material that shows how the totality of these internal and external programs are inter-related and coordinated.

The internal program material should include:

- organizational descriptive material,

- narrative descriptions of each program to be reviewed, and
- descriptive material of each work unit in the program.

It would also prove useful to include bibliometric output indicators for each program, with interpretive analytical material. This could include refereed papers, patents, awards and honors, presentations, etc.

Specifically, internal program background material should include the following administrative and technical information:

- Structural chart of the agency showing how the organization under review fits into agency structure.
- Structural chart of organization, showing programs (including funding) and personnel (including background and expertise) associated with each program.
- Definitions of different generic types of programs that will be presented during the review.
- Administrative material (agenda, reimbursement, conflict-of-interest forms, proprietary protection forms, etc.).
- Two page overview of each program being reviewed in detail (e.g., weapons technology), including:

    - program objective,
    - program thrusts (e.g., aerodynamics, ordnance, guidance and control, etc.),
    - investment allocation among thrusts (three year trends),
    - milestones where appropriate, and
    - progress made toward achieving these milestones.

- Two page overview of each program thrust, including:

    - thrust objective,
    - short descriptions of each technical sub-thrust (e.g., energetic propellants, combustion instability, propellant safety) pursued under the thrust, as well as
    - investment allocations among sub-thrusts.

Total program and thrust descriptive material should not exceed twenty pages. It would be useful to include narrative material on related external programs in other agencies and industry, including descriptions of papers and other output material from these programs, as well as narrative descriptions of ongoing programs. Choice of material sent to reviewers should be very selective, since an excessive amount will go unread. However, it would be useful to include hindsight-type results of research that was funded years ago in the technical area under review, and which recently have come to fruition in a system or commercial technology.

It would also be valuable if roadmaps (Kostoff, 1997d, 2001a) were provided as background material (i.e., visual depictions of the structural relationships among the program components and the mission objectives). These roadmaps provide the global context in which the program is being performed. Retrospective roadmap components depict the program manager's awareness of the breadth and depth of the intellectual heritage of the program being reviewed. Present roadmap components reflect the program manager's awareness of the wide range of S&T areas available to complement his/her program, and the degree of coordination and leveraging in which the program is involved. Prospective roadmap components provide indication of the program manager's vision and willingness to take risks, and his/her intrinsic understanding of how results from other S&T programs could be exploited to enhance and expand the potential of the program. A certain amount of time and reflection is required on the part of the reviewer to understand and to fully appreciate the implications of a well-prepared, comprehensive roadmap. As a result, roadmaps should be sent to reviewers well in advance of the actual review date.

4. Conducting the review

Once the reviewers are assembled, they should be provided with a document containing hard copies of the viewgraphs to be presented, as well as documented evidence of program accomplishments. These accomplishments should include bibliometric information (papers and reports published, conference proceedings, books, awards, etc.), and write-ups of significant accomplishments. Each accomplishment write-up should describe:

- the actual scientific or technological accomplishment,
- what impact it has had, or will have, on

    - other science or technology initiatives,
    - the agency and its national mission, and
    - the performer and performing organization.

The presentations should then occur in the sequence described in section 3.2.3.2. Briefly, a senior agency representative should welcome the reviewers and audience, and describe the purpose of the review from the agency's perspective. The review manager then provides the details of the organization's structure, the types of reviews within the agency, and the integration of the present review with the other reviews and with the total organization's management processes. The review manager also describes the detailed steps of the evaluation process, including the meeting agenda, and presents all the administrative details and procedures to be followed. The head of the organizational unit describes the mission and programs of the unit, and how the program to be reviewed integrates with the remainder of the unit. These presentations constitute the introductory material for the total audience. The program manager then describes the larger context in which the program operates, the structure and contents of the program, and the investment strategy that guides the specific program element allocations. Approximately 1/3 of the presentation period should be devoted to questions and answers.

After the program manager's presentation, time is allotted for written evaluation before proceeding to the next presenter. There is a school of thought that written evaluations should only be performed after a group of presentations rather than after each presentation. This would allow for each presentation to be evaluated in the context of the other presentations, both relative to individual presentations and to the larger collective body of presentations. However, the author has found that an element of spontaneity and freshness is lost by not performing evaluations directly after each presentation. The integrative aspect can be incorporated into the review by allowing for some reflective time, after the day's presentations have been completed, for modifying the written comments, if desired. The executive session at day's end allows for further integration through discussion.

Each of the technical managers then describes his/her S&T sub-category within the program. Again, approximately 1/3 of the presentation time is devoted to questions and answers (Q&A). After each of these presentations, time is allotted for written evaluation before proceeding to the next presenter.

At the end of each presentation day, about one to two hours should be devoted to an executive session, in which the reviewers and review manager meet to discuss each presentation. At the end of the executive session of the final presentation day, all the written evaluation forms are collected. The importance of the verbal (and written) comments made by the discussants depends not only on their intrinsic merit, but on the context in which they are made. It is extremely valuable to have a separate technically knowledgeable observer present throughout the review, who can discuss any contextual issue with the review manager or chairman after the discussions have concluded. This allows key issues to be framed within their proper context in the final report, and allows the credibility of the report to be raised substantially among the sophisticated readers.

5. Post-review actions

After the actual review meetings have been completed, all the information must be assembled, analyzed, and reported. Then actions following the report recommendations must be taken, and the responses to those actions tracked and analyzed. The detailed steps follow.

5.1. Integrating additional comments

Any additional comments about the review, either from the reviewers, the external audience, or senior management should be considered and integrated into the review report, where appropriate. For the reviewers in particular, they have had a chance to integrate all aspects of the review and can provide a cohesive narrative of their views on the program. Either review type, independent panel or individual external reviewer, should insure that this avenue for additional information remains open, not to be arbitrarily closed for some artificial expediency.

5.2. Writing a final report

There should be two forms of the final report, a long version and a short version. The long version should include all the written material that was generated during the course of the review. It provides an archival record of exactly what was done during the review. This report version would include:

- the initial review charter,
- invitation letters,
- background material,
- completed evaluation forms with reviewer identification deleted,
- other reviewer/audience input, and
- the final report write-up.

The short version would summarize the process details, and would focus on reviewer comments and other significant inputs, conclusions, and recommendations. The final report should include the viewpoints of all the reviewers, with appropriate weightings given for judgment and expertise of specific contributors. Dissenting viewpoints should be identified. Based on the diverse inputs, the report author should specify conclusions on the health of the program, and recommendations for action in modifying the program, if required.

5.3. Assigning action items

Under GPRA, there will be at least two clients for the report, internal management, and the Federal government oversight organization. If internal management accepts the conclusions and recommendations of the report, action items should be assigned to the appropriate personnel for responding to problems identified in the report. There are many types of responses possible (e.g., a corrective action, or a rebuttal disagreeing with the conclusion and recommendations). Maximum flexibility and leeway should be given to the program manager for the initial response.

5.4. Evaluating response to action items

Each action item should have a deadline for response. After the deadline, the response should be evaluated, and appropriate follow-up action taken. These action items, responses, and follow-up actions should be presented at the introduction of the next annual review. This provides evidence to the reviewers that their input has impact on the program, and will motivate them to participate in the review process further.

APPENDIX - Executive Summary/Peer-Review Principles (Modified)

The Government Performance and Results Act of 1993 (GPRA, 1993) requires federal agencies to develop strategic plans, annual performance plans, and performance measures to gauge progress in achieving their planned targets. In a companion paper in Science (Kostoff, 1997b), it is recommended that peer review be used as the dominant metric when GPRA is applied to basic research. However, for research program peer review to be used effectively and efficiently for GPRA, it must be understood, developed, and standardized well beyond its present status. In addition, program peer review should be integrated seamlessly into the organization's business operations.evaluation processes in general, and peer review processes in particular. It should not be incorporated in the management tools as an afterthought, as is the case in practice today, but rather should be part of the organization's front-end design. This allows optimal matching among data generating, gathering, and review requirements, and helps avoid the present procedure of force fitting evaluation criteria and processes to whatever data is produced from non-evaluation requirements. This paper focuses on the underlying principles that are necessary for a high quality peer review. While the paper is targeted toward research program peer review, most of the principles are applicable to multiple types of peer review. The author's experience, based on examining the peer review literature, conducting many peer review experiments (e.g., Kostoff, 1988), and managing hundreds of peer reviews, leads to the following conclusions about the factors critical to high-quality peer review (Kostoff, 1995, 1997a, 2001b), whether it be of proposals, programs, procedures, manuscripts, faculty performance, or research dissertations.

1) Senior Management Commitment

The most important factor in a high-quality S&T evaluation is the serious commitment to high-quality S&T evaluations of the evaluating organization's most senior management with evaluation decision authority, and the associated emplacement of rewards and incentives to encourage such evaluations. Incorporated in senior management's commitment to quality evaluations is the assurance that a credible need for the evaluation exists, as well as a strong desire that the evaluation be structured to address that need as directly and completely as possible.

2) Evaluation Manager Motivation

The second most important factor is the operational evaluation manager's motivation to perform a technically credible evaluation. The manager:

a) sets the boundary conditions and constraints on the evaluation's scope;

b) selects the final specific evaluation techniques used;

c) selects the methodologies for how these techniques will be combined/ integrated/ interpreted, and

d) selects the experts who will perform the interpretation of the data output from these techniques.

In particular, if the evaluation manager does notfollow, either consciously or subconsciously, the highest standards in selecting these experts, the evaluation's final conclusions could be substantially determined even before the evaluation process begins. Experts are required for all the evaluation processes considered (peer review, retrospective studies, metrics, economic studies, roadmaps, data mining, and text mining), and this conclusion about expert selection transcends any of these specific applications.

3) The third most important factor is the transmission of a clear and unambiguous statement of the review’s objectives (and conduct) and potential impact/ consequences to all participants. This statement should occur at the very beginning of the review process.

4) Competency of Technical Evaluators

The fourth most important factor is the role, objectivity, and competency of technical experts in any S&T evaluation. While the requirements for experts in peer review, retrospective studies, roadmaps, and text mining are somewhat obvious, there are equally compelling reasons for using experts in metrics-based evaluations. Metrics should not be used as a stand-alone diagnostic instrument (Kostoff, 1997b). Analogous to a medical exam, even quantitative metrics results from suites of instruments require expert interpretation to be placed into proper context and gain credibility. The metrics results should contribute, and be subordinate, to an effective peer review of the technical area being examined.

Thus, this fourth critical factor consists of the evaluation experts' competence and objectivity. Each expert should be technically competent in his subject area, and the competence of the total evaluation team should cover the multiple S&T areas critically related to the science or technology area of present interest. In addition, the team's focus should not be limited to disciplines related only to the present technology area (that tends to reinforce the status quo and provide conclusions along very narrow lines). It should be broadened to disciplines and technologies that have the potential to impact the overall evaluation's highest-level objectives (that would be more likely to provide equitable consideration to revolutionary new paradigms).

5) Selection of Evaluation Criteria

The fifth most important factor is selection of evaluation criteria (Delcomyn, 1991; Sutherland, 1993; Weinberg, 1989). These criteria will depend on the:

- interests of the audience for the evaluation,
- nature of the benefits and impacts,
- availability and quality of the underlying data,
- accuracy and quality of results desired,
- complementary criteria available and suites of diagnostic techniques desired for the complete analysis,
- status of algorithms and analysis techniques, and
- capabilities of the evaluation team.

For evaluating basic research proposals, the three main criteria are research merit, research approach, and team quality (DOE, 1982; Kostoff, 1992, 1997a). For research sponsored by a mission-oriented organization, a fourth criterion related to mission relevance is useful. To ensure that this mission relevance criterion does not filter out the more basic research oriented proposals, a very liberal interpretation of mission relevance is necessary. For basic research, a nearer-term relevance criterion, such as transition or utility, correlates better with overall proposal quality score than does a longer-term criterion (Kostoff, 1992). Use of a fifth criterion for overall research quality is essential, and makes it possible to incorporate the effects of unlisted criteria that the reviewer feels is important for considering a specific proposal. For example, reviewers might feel that an agency proposal is more appropriate for sponsorship by industry than by government. In this case, the proposal could receive a low overall rating, even though the listed component technical criteria were rated very high.

6) Relevance of Evaluation Criteria to Future Action
A factor of equal importance to evaluation criteria selection is one that has been violated in almost every metrics briefing the author has attended spanning many government agencies, industrial organizations, and academic institutions. In general, this factor tends to be violated for the evaluation criteria used in any of the evaluation approaches under the decision aids umbrella. The factor will be stated in terms of a metrics-based evaluation, but it should be considered as applicable to all evaluation techniques.

EVERY S&T METRIC, AND ASSOCIATED DATA, PRESENTED IN A STUDY OR BRIEFING SHOULD HAVE A DECISION FOCUS; IT SHOULD CONTRIBUTE TO THE ANSWER OF A QUESTION WHICH IN TURN WOULD BE THE BASIS OF A RECOMMENDATION FOR FUTURE ACTION.

Metrics and associated data that do not perform this function become an end in themselves, offer no insight to the central focus of the study or briefing, and provide no contribution to decision-making. They dilute the theme of the study, and, over time, tend to devalue the worth of metrics in credible S&T evaluations. Because of:

1) the political popularity and subsequent proliferation of S&T metrics;
2) the widespread availability of data; and
3) the ease with which this data can be electronically gathered/ aggregated/ displayed,

most S&T metrics briefings and studies are immersed in data geared to impress rather than inform. While metrics studies provide the most obvious examples, this conclusion can be easily generalized to any of the evaluation methods.

7) Reliability of Evaluation
Another factor of equal importance is reliability or repeatability. To what degree would an S&T evaluation be replicated if a completely different team were involved in selection, analysis, and interpretation of the basic data? If each evaluation team were to generate different evaluation criteria, and in particular, generate far different

interpretations of these criteria for the same topic, then what meaning or credibility or value can be assigned to any S&T evaluation (Cole, 1981)? To minimize repeatability problems, a diverse and representative segment of the overall competent technical community should be involved in the construction and execution of the evaluation.

8) Evaluation Integration

A fourth factor of equal importance is the seamless integration of evaluation processes in general into the organization's business operations. Evaluation processes should not be incorporated in the management tools as an afterthought, as is the case in practice today, but should be part of the organization's front-end design. This allows optimal matching between data generating/ gathering and evaluation requirements, not the present procedure of force fitting evaluation criteria and processes to whatever data is produced from non-evaluation requirements.

9) Global Data Awareness

A fifth factor of equal importance is data awareness (Kostoff, 1999a, 2000a, 2000b, 2001d, 2003a, 2003c). In all of the decision aids, placement of the technology of interest in the larger context of technology development and availability world-wide is an absolute necessity. This tends to be a central deficiency of most management decision aids. Lack of S&T documentation, inaccessibility of S&T that is documented, inability to retrieve S&T documents due to poor retrieval methods, inability to extract information from large retrievals, and general lack of interest and will in global data awareness, mitigate against attaining comprehensive global data awareness.

10) Normalization across Technical Disciplines

For evaluations that will be used as a basis for comparison of S&T programs or projects, the next most important factor is normalization and standardization across different S&T areas. For S&T areas that have some similarity, use of common experts (on the evaluation teams) with broad backgrounds that overlap the disciplines can provide some degree of standardization (Kostoff, 1988, 1997a). For very disparate S&T areas, some allowances need to be made for the relative strategic value of each discipline to the organization, and arbitrary corrections applied for benefit estimation differences and biases. Even in this case of disparate disciplines, some normalization is possible by having some common team members with broad backgrounds contributing to the evaluations for diverse programs and projects (Van den Beemt, 1997). However, normalization of the criteria interpretation for each science or technology area's unique characteristics is a fundamental requirement.

Because credible normalization requires substantial time and judgement, it tends to be an operational area where quality is sacrificed for expediency.

11) Reviewer Anonymity

A factor of equal importance to normalization is secrecy: reviewer anonymity and reviewee non-anonymity (Altura, 1990; Clayson, 1995; Gresty, 1995; Neetens, 1995). If honest and frank viewpoints on the intrinsic quality of the research under review are desired, the reviewer must remain anonymous to all but the review manager. Rewards are few for a reviewer making strong negative statements about a proposal (or research paper or program), and resulting retributions and resentments to the reviewer may far outweigh the intrinsic benefits to science of honest and forthright judgment statements.

"Blind reviewing," the withholding of the reviewee's name and affiliation from the reviewer, has been used for the noble purposes of providing fairer reviews of work by unknown researchers or by researchers from less prestigious institutions, and to eliminate bias based on personal characteristics such as gender (Ceci, 1984; Laband, 1994; Cox, 1993; Nylenna, 1994). However, studies of proposed and existing research evaluations have shown that team quality was the most important variable in determining overall project quality (DOE, 1982). Removing the identity of the reviewee from the research under review is akin to solving an equation after eliminating the dominant term. As a result, rather than eliminate the key variable of researcher identity, it may be more important to select additional reviewers who will broaden the review group's perspective and address the "right job" aspects of the research project. This will help insure that outmoded, albeit frequently cited, research is not promulgated in perpetuity, and that fresh perspectives of new paradigms will receive the attention they deserve.

12) Cost of S&T Evaluations

The next critical factor for quality S&T evaluations is cost (ASTEC, 1991; Buechner, 1974; Hensley, 1980; Kostoff, 1995, 1997a). The true total costs of peer review can be considerable, but tend to be ignored or understated in most reported cases. For high quality peer reviews, where sufficient expertise is represented on the review group, total real costs will dominate direct costs (Kostoff, 1995, 1997a). The major contributor to total costs is the time of all the individuals involved in executing the review, including staff, reviewer, and presenter time. If a substantial audience is in attendance, then audience time should be included in review costs. With high quality performers and reviewers, time costs are high, and the total review costs can be non-negligible. For sponsor environments where a large number of proposals are rejected, and where multiple proposals to different sponsors are the norm, peer review costs per funded proposal increase dramatically in proportion to the ratio of proposals reviewed to proposals funded. Accurate cost analyses should not be neglected in designing a high quality proposal, manuscript, or program peer-review process.

13) Maintenance of High Ethical Standards
The final critical factor, and perhaps the foundational factor, in any high quality S&T evaluation is the maintenance of high ethical standards throughout the process. There is a plethora of potential ethical issues (Fielder, 1995; Goodstein, 1995; Gupta, 1996; Keown, 1996; Moran, 1992), including technical fraud, technical misconduct, betraying confidential information, and unduly profiting from access to privileged information. This stems from an inherent bias/ conflict of interest in the process when real experts are desired to participate in every aspect of an S&T evaluation. The evaluation managers need to be vigilant for undue signs of distortion aimed at personal gain.

BIBLIOGRAPHY

Altura, B.T. 1990. Is Anonymous Peer-Review The Best Way To Review And Accept Manuscripts? In: Magnesium And Trace Elements. 9:117-118.

Armstrong, J.S. 1997. Why Conduct Journal Peer Review: Quality Control, Fairness, Or Innovation. Sci Engineer Ethics, 3:1.

ASTEC. 1991. Funding The Fabric - Should Commonwealth Government Competitive Research Granting Schemes Contribute More To Research Infrastructure Costs? Australian Government Publishing Service, Canberra, Australia.

Brown, E.A. 1996. Conforming The Government R&D Function With The Requirements Of The Government Performance And Results Act: Planning The Unplannable? Measuring The Unmeasurable? Scientometrics 36:3.

Buechner, Q. 1974. Proposal Costs. J Soc Res Adminis 5:47-50.
Ceci, S.J. And D. Peters. 1984. How Blind Is Blind Review? Amer Psychol 39:1491-1494.

Chubin, D.E. And E.J. Hackett. 1990. Peerless Science: Peer Review And U.S. Science Policy. State University Of New York Press, Albany, New York.

Clayson, D.B. 1995. Anonymity In Peer-Review - Time For A Change - Comment. Regulatory Toxicol Pharmacol 22:101-101.

Cole, J.R. & Cole, S. "Peer Review in the National Science Foundation: PhaseTwo of a Study". Washington, DC. National Academy Press. 1981.

Cox, D., L. Gleser, N. Perlman, N. Reid, And K. Roeder. 1993. Report Of The Ad-Hoc Committee On Double-Blind-Refereeing. Statist Sci 8:310-317.

Delcomyn, F. 1991. Peer-Review - Explicit Criteria And Training Can Help. Behavior Brain Sci 14:144-144.

Department Of Energy. 1982. An Assessment Of The Basic Energy Sciences Program. Office Of Energy Research, Office Of Program Analysis. Report No. DOE/ER-0123 (March 1982).

Fielder, J.H. 1995. Disposable Doctors - Incentives To Abuse Physician Peer-Review. J Clinic Ethics. 6:327-332.

Goodstein, D. 1995. Ethics And Peer-Review - Commentary. Stem Cells 13:574-574.

GPRA. 1993. Government Performance And Results Act Of 1993. PL 103-62.

Gresty, M.A. 1995. Peer-Review And Anonymity. Neurol-Ophthamol 15:281-282.

Gupta,V.K. 1996. Should Intellectual Property Be Disseminated By Forwarding Rejected Letters Without Permission? J Med Ethics 22:243-244.

Hensley, O., B. Gulley, And J. Eddleman. 1980. Evaluating Development Costs For A Proposal To A Federal Agency. J Soc Res Adminis 12:35-39.

Keown, D. 1996. The Journal Of Buddhist Ethics - An Online Journal. Learned Publish 9:141-145.

Kostoff, R.N. 1988. Evaluation Of Proposed And Existing Accelerated Research Programs By The Office Of Naval Research. IEEE Trans Engineer Manage 35:4 Nov.

Kostoff, R.N. 1992. Research Impact Assessment. Proceedings: Third International Conference On Management Of Technology, Miami, FL (February 17-21). (Larger Text Available From Author.)

Kostoff, R.N. 1995. Federal Research Impact Assessment: Axioms, Approaches, Applications. Scientometrics 34:2.

Kostoff, R.N. 1997a. The Handbook Of Research Impact Assessment (7th Ed.). DTIC Report Number ADA-296021. (See Also Http://Www.Dtic.Mil/Dtic/Kostoff/Index.Html)

Kostoff, R.N. 1997b. Peer Review: The Appropriate GPRA Metric For Research. Science 277:651-652.

Kostoff, R.N. 1997c. Research Program Peer Review: Principles, Practices, Protocols. (Http://Www.Dtic.Mil/Dtic/Kostoff/Index.Html).

Kostoff, R.N. 1997d. Science And Technology Roadmaps. (Http://Www.Dtic.Mil/Dtic/Kostoff/Index.Html).

Kostoff, R.N. 1997e. Science And Technology Innovation. (Http://Www.Dtic.Mil/Dtic/Kostoff/Index.Html).

Kostoff, R. N. 1997f. The Principles And Practices Of Peer Review, In: Stamps, A. E., (Ed.), Science And Engineering Ethics, Special Issue On Peer Review, 3:1.

Kostoff, R. N. 1997g. Use And Misuse Of Metrics In Research Evaluation, Science And Engineering Ethics, 3:2.

Kostoff, R. N., And Geisler, E. 1999a. Strategic Management And Implementation Of Textual Data Mining In Government Organizations. Technology Analysis And Strategic Management. 11:4.

Kostoff, R. N. 1999b. Science And Technology Innovation. Technovation. 19:10. 593-604. October 1999.

Kostoff, R. N. 2000a. Science And Technology Text Mining. Keynote Presentation/ Proceedings. TTCP/ ITWP Workshop. Farnborough, UK. 12 October.

Kostoff, R. N. 2000b. Implementation Of Textual Data Mining In Government Organizations. Proceedings: Federal Data Mining Symposium And Exposition, 28-29 March.

Kostoff, R. N., And Schaller, R. R. 2001a. Science And Technology Roadmaps. IEEE Transactions On Engineering Management. 48:2. 132-143. May.

Kostoff, R. N., Miller, R., Tshiteya, R. 2001b. Advanced Technology Development Program Review – A US Department Of The Navy Case Study. R&D Management. 31:3. 287-298. July.

Kostoff, R. N., And Demarco, R. A. 2001c. Science And Technology Text Mining. Analytical Chemistry. 73:13. 370-378A. 1 July.

Kostoff, R. N. 2001d. The Extraction Of Useful Information From The Biomedical Literature". Academic Medicine. 76:12. December.

Kostoff, R. N. 2003a. Text Mining For Global Technology Watch. Encyclopedia Of Library And Information Science. In Press.

Kostoff, R.N. 2003b. Role Of Technical Literature In Science And Technology Development. Journal Of Information Science. In Press.

Kostoff, R. N. 2003c. Data – A Strategic Resource For National Security. Academic And Applied Research In Military Science. In Press.

Kostoff, R. N. 2003d. Disruptive Technology Roadmaps. Technology Forecasting And Social Change. In Press.

Laband, D.N. 1994. A Citation Analysis Of The Impact Of Blinded Peer-Review. J. Amer Med Assoc 272:2.

Moran, G. 1992. Ethical Questions About Peer-Review. J Med Ethics 18:160-160.

Neetens, A. 1995. Should Peer Reviewers Shed The Mask Of Anonymity. Neuro-Ophthalmol 15:109-109.

Nylenna, M., P. Riis, And Y. Karlsson. 1994. Multiple Blinded Reviews Of The 2 Manuscripts - Effects Of Referee Characteristics And Publication Language. J Amer Med Assoc 272:149-151.

Sutherland, H.J., E.M. Meslin, R. Dacunha, And J.E. Till. 1993. Judging Clinical Research Questions - What Criteria Are Used. Social Sci Med 37:1427-1430.

Van Den Beemt, F.C.H.D. And C. Le Pair. 1991. Grading The Grain: Consistent Evaluation Of Research Proposals. Res Evalu 1:1.

Van Den Beemt, F.C.H.D. 1997. The Right Mix: Review By Peers As Well As By Highly Qualified Persons (Non-Peers). Australian Res Council Commissioned Report: "Peer Review Process" No. 54. Pp. 153-164.

Weinberg, A.M. 1989. Criteria For Evaluation, A Generation Later. In: Ciba Foundation (Ed.). The Evaluation Of Scientific Research (John Wiley & Sons). Pp. 3-12.